Insertion site-based polymorphism: A Swiss army knife for wheat genomics

Etienne Paux[1], Lifeng Gao[1], Sébastien Faure[1], Frédéric Choulet[1], Delphine Roger[2], Karine Chevalier[1], Cyrille Saintenac[1], François Balfourier[1], Karine Paux[1], Mehmet Cakir[3], Dominique Brunel[4], Marie-Christine Le Paslier[4], Tamar Krugman[5], Béatrice Gandon[2], Eviatar Nevo[5], Michel Bernard[1], Pierre Sourdille[1], Catherine Feuillet[1]

[1] UMR GDEC, INRA, Clermont-Ferrand, France. [2] Limagrain Verneuil Holding, Riom, France. [3] State Agricultural Biotechnology Center, Murdoch University, Murdoch, Australia. [4] UR EPGV, INRA, Evry, France. [5] Institute of Evolution, University of Haifa, Israel.

INTRODUCTION

Transposable elements (TEs) are prevalent in the genomes of all plants. They are ubiquitous, in high-copy number, evenly distributed in the genome, in both hetero- and euchromatin, and show insertional polymorphism both between and within species[1,2]. These genetic properties have recently allowed the development of several TE-based molecular marker types, such as S-SAP, IRAP, REMAP and RBIP[3,4]. These molecular markers have successfully been used to establish phylogenies, study biodiversity and generate linkage maps for agronomically important traits in several species such as barley, pea, rice and tobacco[4-6]. In wheat, TEs account for more than 70% of the genome[2,7] and play a major role in the structure and evolution of the genome. It is likely that TEs have driven wheat genome evolution in diverse ways, including genome expansion and contraction, segmental duplication, and exon shuffling. It has been proposed that TE-induced genomic rearrangements tend to promote both cytological and genetic diploidization of polyploid genomes[8,9]. Therefore, TE-based molecular markers represent ideal tools to study the structure and evolution of the hexaploid wheat genome.

In the framework of a BAC-end sequencing project on wheat chromosome 3B, we have recently demonstrated the potential of small genomic sequences for developing genome-specific TE-based molecular markers. We established a method, called Insertion Site-Based Polymorphism (ISBP) that exploits knowledge of the sequence flanking a TE to amplify by PCR a fragment spanning the junction between the TE and the flanking sequence[2]. Several hundreds of ISBP markers evenly distributed along the chromosome 3B of bread wheat and representative of all kind of junctions (various TE families in both repetitive and low copy DNA, either coding or non-coding) have been defined.

Here, we report the development of a bioinformatics tool for the automated design of ISBP markers as well as the implementation of several genotyping techniques. We also demonstrate the usefulness of ISBP markers as a new tool for wheat genomic studies.

AUTOMATED DESIGN OF ISBP MARKERS FROM GENOMIC SEQUENCES

One of the advantages of the ISBP markers is their straightforward design from short genomic sequences, as previously demonstrated using BAC-end sequences[2]. The recent progress in next generation sequencing technologies such as the Roche 454 Genome Sequencer FLX opens new perspectives for the high throughput development of ISBP markers.

To fully benefit from this major breakthrough in genome sequencing, we developed software for the automated design of ISBP markers. This program, called ‘IsbpFinder.pl’, uses annotation results generated by the REPEATMASKER program[10] with TREP[11] as a custom library to detect the junction between TEs and designs primers to amplify a genomic fragment spanning this junction (Figure 1).

To assess the efficiency of this software for ISBP design, we used two different datasets. The first one comprised approximately 50,000 BAC-end sequences (BES) originating from chromosome 3AS- and chromosome 3B-specific BAC libraries. Useful junctions between TEs were identified in roughly 5% of the BES leading to the design of about 2500 putative ISBPs. The second set corresponded to a 3.2-Mb BAC contig sequence from chromosome 3B (Choulet et al., unpublished data). Using IsbpFinder.pl, we were able to design about 1000 putative ISBP markers, corresponding to an average of one marker per 3 kb. The design was subsequently validated on a subset of markers. The success rate of ISBP design (*i.e.* the probability for a predicted ISBP to correspond to a single genomic locus) was about 70%.

Considering that ~14 out of the 17 Gb wheat genome is comprised of repetitive DNA, ISBPs represent an almost infinite source of polymorphism in wheat. Indeed, based on a density of one ISBP per 3 kb and a success rate of 70%, we can estimate the total number of ISBPs to be close to 4 million in the hexaploid wheat genome, the largest source of polymorphism ever produced in wheat. Moreover, because insertion sites are fairly unique in the genome, ISBP markers are mostly genome specific. This potential amount of genome-specific markers is likely to allow saturation of genetic maps and subsequently unlock many doors leading to efficient genetic diversity studies, recombination and linkage disequilibrium analyses, association mapping, fine mapping and cloning of QTLs, as well as marker assisted-selection.

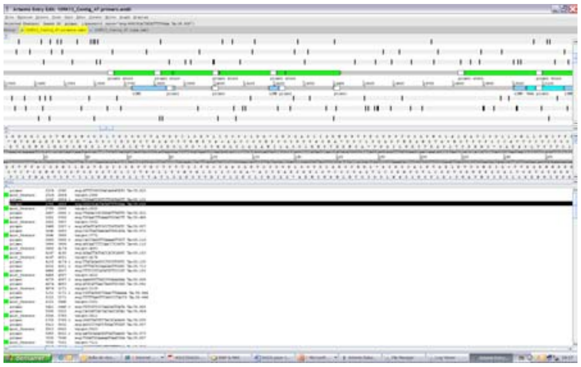

Figure 1. Automated design of ISBP from a BAC sequence using IsbpFinder. The EMBL-formatted output can be visualized using Artemis software.

A WIDE RANGE OF DETECTION TECHNIQUES

ISBP markers have been initially implemented on classical agarose gel electrophoresis[2]. This technique allows for simple detection of amplicons and can therefore be used to visualize presence / absence polymorphism as well as length polymorphism assuming that the size difference between two alleles is large enough to be visualized on a gel. However, it is limited by its low throughput and low resolution. To overcome these limitations, we have implemented a range of other detection techniques (Figure 2).

Melting curve analysis. The presence or absence of ISBP amplicons as well as differences in sequence length or composition can be scored using melting curve analysis as each double-stranded DNA has its own specific melting temperature (Tm), which is determined by DNA length and GC content. This technique allows for a high-throughput and cost-effective genotyping of ISBPs but is limited in terms of resolution and is not suitable for heterozygous detection.

Fluorescent PCR and capillary electrophoresis. PCR performed using fluorescent-labelled primers allows for high-throughput estimate of amplicon length on capillary sequencers thereby leading to the rapid detection of product length polymorphisms as small as 2 bp and up to several hundreds of nucleotides. However, differences in sequence composition cannot be detected.

Temperature gradient capillary electrophoresis (TGCE). This technique can be used to detect ISBP polymorphism between two genotypes, without prior knowledge of the single nucleotide polymorphism (SNP)[12]. However, as it allows only for comparison between two genotypes, TGCE is mainly limited to the genetic mapping of ISBPs and cannot be used for diversity studies.

Allele-specific PCR (AS-PCR). This technique is based on the selective amplification of one of the ISBP alleles to detect SNPs[13]. Selective amplification is achieved by designing a primer such that the primer will match/mismatch one of the alleles at the 3'-end of the primer. The combination of the two allele-unspecific ISBP primers with one allele-specific primer allows for heterozygous genotyping and thus mapping in F2 populations. However, preliminary sequencing of the ISBP is needed as AS-PCR requires prior knowledge of the sequence polymorphism.

SNaPshot. Similarly to AS-PCR, this primer extension-based technique[14] can be used for efficient ISBP genotyping including heterozygous lines. The combination of fluorescent labelling with capillary electrophoresis allows for marker multiplexing and high-throughput genotyping. However, due to the repetitive nature of ISBPs, a preliminary amplification is required to specifically target a single locus.

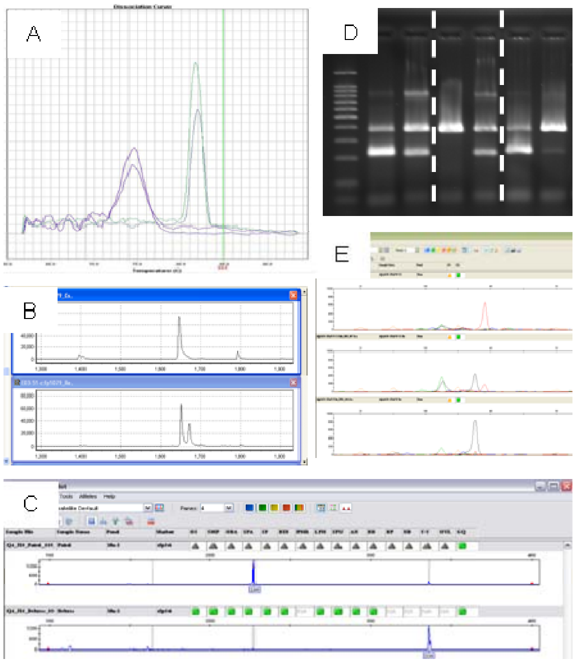

Figure 2. Detection techniques for ISBP genotyping. (A) Melting curve analysis; (B) Fluorescent PCR and capillary electrophoresis; (C) Temperature gradient capillary electrophoresis; (D) Allele-specific PCR; (E) SNaPshot.

PHYLOGENY OF WHEAT AND WILD RELATIVES

Bread wheat is an allohexaploid whose ABD-genome derived from spontaneous hybridization events between three homoeologous diploid genomes. While the diploid donors of the A- and D-genomes have been identified quite confidently as being *Triticum urartu* and *Aegilops tauschii*, respectively, the origin of the B-genome remains controversial. Nevertheless, the B-genome is supposed to derive from species related to the S-genome of *Aegilops* section *Sitopsis* that includes *Ae. speltoides*, *Ae. bicornis*, *Ae. longissima*, *Ae. searsii*, and *Ae. sharonensis*[15]. Another question is the origin of the G-genome of *Triticum timopheevi*, which is closely related to the B-genome and frequently reported to originate from *Ae. speltoides*[16].

To address these questions, we used ISBPs distributed along chromosome 3B to genotype roughly 400 accessions of wheat and wild relatives: *Aegilops* of the *Sitopsis* section (S-genome), *T. turgidum* (AB-genome), hexaploid wheat (ABD-genome) and *T. timopheevii* (AG-genome).

Analysis of ISBP allelic diversity among these accessions indicated that more than 75% of the genetic variation is found among populations. The highest diversity was observed in wild emmer wheat (*T. dicoccoides). Ae. speltoides* appeared to be phylogenetically distinct from the *Sitopsis* section and closer to the G-genome than to the B-genome. This supports the hypothesis of a distinct origin of the B- and G-genome. Moreover, the results suggest that the B-genomes of hexaploid and domesticated tetraploid wheat likely originated from *T. dicoccum*. Finally, the phylogenetic proximity of domesticated hexaploid and tetraploid wheats strongly suggests an effect of domestication on B-genome diversity.

EFFECTS OF THE ENVIRONMENTAL CONDITIONS ON TE-INDUCED GENOMIC VARIABILITY

Although most of the TEs are quiescent in the plant genomes, they can be activated in response to biotic and abiotic stresses[17]. For example, amplification and losses of the BARE-1 family of LTR-retrotransposons was shown to generate genomic diversity in plants under the influence of environmental factors[18]. ISBP markers provide an ideal tool for evaluating to what extent TE-induced genomic variability can be generated under environmental conditions in wheat.

To this aim, we genotyped a collection of wild emmer wheat (*Triticum dicoccoides*) from different populations representing regional patterns as well as contrasting microsites (differences in soil, vegetation, sun exposure…) in Israel with ISBP markers of chromosome 3B. ISBPs allowed the clear discrimination between almost all populations. Interestingly, we found correlations between several environmental factors and gene diversity strongly suggesting an impact of environmental conditions on TE transposition and subsequent TE-induced genomic variability.

ISBP MARKERS AS TOOLS FOR MARKER-ASSISTED SELECTION

In wheat, the widespread application of marker-assisted selection is currently hampered by the lack of high-throughput markers. ISBPs represent an almost infinite source of polymorphism and have therefore the potential to overcome this limitation.

To assess the usefulness of ISBPs for marker-assisted selection, ISBPs from chromosome 3B were used to genotype European and Australian elite wheat varieties. In total, 60% of the markers were polymorphic with the number of alleles ranging from 2 to 6. Using melting curve analysis, we were able to discriminate between Australian and European lines. Together, these results demonstrate the potential of ISBP markers in wheat breeding programs.

REFERENCES

1. Kumar A et al. (1997) *Genetica*, 100, 205-217.
2. Paux E et al. (2006) *Plant J.*, 48, 463-474.
3. Kumar A and Hirochika H (2001) *Trends Plant Sci.*, 6, 127-134.
4. Schulman AH et al. (2004) *Methods Mol. Biol.*, 260, 145-173.
5. Kalendar R et al. (1999) *Theor. Appl. Genet.*, 98, 704-711.
6. Kenward KD et al. (1999) *Theor. Appl. Genet.*, 98, 387-395.
7. Li W et al. (2004) *Plant J.*, 40, 500-511.
8. Levy AA and Feldman M (2002) *Plant Physiol.*, 130, 1587-1593.
9. Feldman M and Levy AA (2005) *Cytogenet. Genome Res.*, 109, 250-258.
10. Smit A et al. (1996). *RepeatMasker*. http://www.repeatmasker.org.
11. Wicker T et al. (2002) *Trends Plant Sci.*, 7, 561-562.
12. Hsia AP et al. (2005) *Theor. Appl. Genet.*, 111, 218-225.
13. Okayama H et al. (1989) *J. Lab. Clin. Med.*, 114, 105-113.
14. Pati N et al. (2004) *J. Biochem. Biophys. Methods*, 60, 1-12.
15. Petersen G et al. (2006) *Mol. Phylogenet. Evol.*, 39, 70-82.
16. Kilian B et al. (2007) *Mol. Biol. Evol.*, 24, 217-227.
17. Casacuberta JM and Santiago N (2003) *Gene*, 311, 1-11.
18. Kalendar R et al. (2000) *Proc. Natl Acad. Sci. U.S.A.*, 97, 6603-6607.